Identifying Problem Localization in Peer-Review Feedback

Wenting Xiong and Diane Litman

University of Pittsburgh

Abstract. In this paper, we use supervised machine learning to automatically identify the problem localization of peer-review feedback. Using five features extracted via Natural Language Processing techniques, the learned model significantly outperforms a standard baseline. Our work suggests that it is feasible for future tutoring systems to generate assessments regarding the use of localization in student peer reviews.

Keywords: peer-review, problem localization, Natural Language Processing.

1 Introduction

There is increasing interest in building systems such as SWoRD[1] to facilitate peer-review practices, which involve students writing essays on certain prompts, reviewing essays for their peers by providing feedback and then revising their previous draft essays based on the peer feedback. However, such systems do not tutor students to write better reviews. A study of a SWoRD corpus [1] shows that the helpfulness of the feedback (in terms of the likelihood of students' revising based on it) is significantly affected by certain feedback features, among which problem localization is most important. While such feedback features were used as mediators in the analysis of feedback helpfulness in [1], we believe that they could also be used as indicators in evaluating feedback quality automatically. As a first step, we focus on predicting problem localization based on the findings noted above, while our long-term goal is to enrich current peer-review systems with an assessment component on student reviewing performance. We illustrate the successful use of supervised machine learning to automatically identify the problem localization for a given piece of feedback based on features obtained using Natural Language Processing (NLP) techniques.

2 Data and Method

This study uses an annotated peer review corpus [1] collected from a college history class. It consists of 874 pieces of feedback expressing criticism accompanied

[1] Scaffolded Writing and Reviewing in the Discipline, http://www.lrdc.pitt.edu/schunn/sword/index.html

V. Aleven, J. Kay, and J. Mostow (Eds.): ITS 2010, Part II, LNCS 6095, pp. 429–431, 2010.

by 24 corresponding essays. The feedback has been segmented at the idea-unit level and coded for problem localization as a binary feature (Kappa=0.69).

We developed four groups of features to capture different perspectives of localized expressions as follows:

Regular expression features: regTag
Simple regular expressions were employed to recognize common phrases of location (e.g., "on page 5", "the section about"). If any regular expression is matched, the binary feature regTag is true.

Domain lexicon features: dwCNT
Using standard statistical NLP techniques provided by NLTK[2] (to extract frequent lexical bigrams from text), a dictionary of domain words was generated automatically from the collection of the 24 essays. We counted those words (dwCNT) contained in each piece of feedback.

Syntactic features: SO_domain, DET_CNT
Besides just counting the domain words, we also extracted information from the syntactic structure of the feedback sentences. We investigated whether there is any domain word between the subject and the object (SO_domain) in any sentence, and also counted demonstrative determiners (this, that, these and those) in the feedback (DET_CNT).

To illustrate how these features were computed, consider the sentence below, which is an idea unit that is coded as "problem localization = true". The regTag is true because one regular expression is matched with "the section of"; dwCNT is 9, because the sentence contains "African" (2), "American", "Americans", "federal", "governments", "civil", "political" and "rights". There is no demonstrative determiner, thus DET_CNT is zero; "African Americans" is between the subject "section" and the object "attention", so SO_domain is true.

> ***Example:*** *The section of the essay on African Americans needs more careful attention to the timing and reasons for the federal governments decision to stop protecting African American civil and political rights.*

Overlapping-window features: windowSize, overlapNum
The three types of features above are based on our intuition about localized expressions, while the following features are derived from an overlapping-window algorithm that was shown to be effective in a similar task – identifying quotation from reference works in primary materials for digital libraries [2]. To match a possible citation in a reference work, it searches for the most likely referred window of words through all possible primary materials. We applied this algorithm for our purpose, and considered the length of the window (windowSize) plus the number of overlapped words in the window (overlapNum).

3 Results

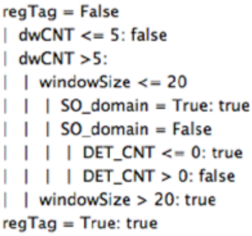
Our binary classifier of problem localization is learned by using the Decision Tree (J48) algorithm provided by WEKA[3] based on the features explained in the

[2] Natural Language Toolkit: http://www.nltk.org/

[3] http://www.cs.waikato.ac.nz/ml/weka/

Metric	Baseline	Learned model	
Accuracy	0.529	0.774	*
Precision	0.279	0.779	*
Recall	0.529	0.773	*
Kappa	0	0.549	*

Fig. 1. Performance of identification of problem localization. * indicates $P < 0.05$.



```
regTag = False
| dwCNT <= 5: false
| dwCNT >5:
| | windowSize <= 20
| | | SO_domain = True: true
| | | SO_domain = False
| | | | DET_CNT <= 0: true
| | | | DET_CNT > 0: false
| | windowSize > 20: true
regTag = True: true
```

Fig. 2. Learned decision tree

previous section. We evaluated our model via 10-fold cross validation and compared its performance against a standard baseline – majority class (always predict "true").

The results presented in Fig 1. show that our model performs significantly better than the baseline. Accuracy is 77.4% (against 52.9%), and both precision and recall are around 77% (against 27.9% and 52.9% respectively). Fig. 2 shows the decision tree based on all the features we investigated. WEKA automatically selects the most powerful features (i.e. regTag, dwCNT, windowSize, SO_domain, DET_CNT) and ignores the less useful ones. The learned model first uses regular expressions to recognize the localized feedback; for feedback whose regTag is false, it then looks at the occurrences of domain words. For domain-word counts greater than 5, the overlapped content between feedback and its targeting essay is then considered, and so on.

4 Conclusion and Future Work

In this paper, we proposed a model for detecting problem localization of peer-review feedback. We found simple NLP techniques (i.e. regular expressions, lexicon dictionaries and text mapping) are effective in our identification task.

In future work, we would like to explore more sophisticated NLP techniques to improve our current model; we would also like to investigate how to generate assessment regarding problem localization based on the noisy output of the model. Furthermore, we hope to incorporate this assessment component into the peer-review system (e.g. SWoRD) so as to provide meaningful feedback for students to enhance their reviewing skills in focused aspects (currently it is just problem localization) from the peer review assignment.

Acknowledgement. We thank LRDC for financial support, M. Nelson and C. Schunn for the corpus, and Rebecca Hwa for the guidance.

References

1. Nelson, M.M., Schunn, C.D.: The nature of feedback: how different types of peer feedback affet writing performance. Instructional Science 37, 375–401 (2009)
2. Ernst-Gerlach, A., Crane, G.: Identifying quotations in reference works and primary materials. In: Christensen-Dalsgaard, B., Castelli, D., Ammitzbøll Jurik, B., Lippincott, J. (eds.) ECDL 2008. LNCS, vol. 5173, pp. 78–87. Springer, Heidelberg (2008)